Downloaded from orbit.dtu.dk on: May 05, 2023

DTU Library

SciCloud: A Scientific Cloud and Management Platform for Smart City Data

Liu, Xiufeng; Nielsen, Per Sieverts; Heller, Alfred; Gianniou, Panagiota

Published in:
2017 28th International Workshop on Database and Expert Systems Applications (DEXA)

Link to article, DOI:
10.1109/DEXA.2017.22

Publication date:
2017

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Liu, X., Nielsen, P. S., Heller, A., & Gianniou, P. (2017). SciCloud: A Scientific Cloud and Management Platform for Smart City Data. In *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)* (pp. 27-31). IEEE. https://doi.org/10.1109/DEXA.2017.22

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

2017 28th International Workshop on Database and Expert Systems Applications

SciCloud: A Scientific Cloud and Management Platform for Smart City Data

Xiufeng Liu and Per Sieverts Nielsen

Department of Management gineering
Technical University of Denmark
{xiuli, pernn}@dtu.dk

Alfred Heller and Panagiota Gianniou

Department of Civil Engineering
Technical University of Denmark
{alfh, pagian}@byg.dtu.dk

***Abstract*—The pervasive use of Internet of Things and smart meter technologies in smart cities increases the complexity of managing the data, due to their sizes, diversity, and privacy issues. This requires an innovate solution to process and manage the data effectively. This paper presents an elastic private scientific cloud, *SciCloud*, to tackle these grand challenges. SciCloud provides on-demand computing resource provisions, a scalable data management platform and an in-place data analytics environment to support the scientific research using smart city data.**

***Keywords*—Cloud, Platform, IoT, Smart city data.**

I. Introduction

Smart cities have been under development globally over the past decade, with a particular focus on the use of information and communication technologies (ICT) to manage urban infrastructure, including building, energy, transport and pollution monitoring [20]. Internet of Things (IoT) is widely deployed to make cities more green, safer and more efficient. Sensors and other smart devices are connected to urban infrastructure to obtain real-time information for decision-making purposes. IoT-based services will continue a substantial development, which is expected to reach 212 billion deployed entities globally by the end of 2020 [11]. In addition, it is clear that smart cities have been revolutionized by cloud-based ICT infrastructures to address the complex IoT services required for urban infrastructure. Cloud-based ICT infrastructures can integrate IoT technologies as needed, and can collect real-time data and the data from other sources, such as operational systems, legacy systems, and applications. According to the recent Forbes survey [21], more than 80 percent of today's organizations are using at least one cloud-based service in their businesses. A typical example is geospatial representation, which is showing an increasing popularity in recent years, such as in the application of urban planning and building modelling [7], [14]. This also illustrates the need for innovative solutions for building and geographic data integration, e.g., integrating the information of all aspects of buildings including construction details, related energy and more [7]. The data can be provided as a service to city governors, urban planners and citizens for decision making.

Moreover, the Big Data trend generates data increasingly complex and large, which makes analysis, archiving and sharing challenging. For example, in our current smart city work, a single building with indoor air-quality sensors can easily generate more than 10,000 data points per minute. In fact, smart cities have many use cases that produce large data sets, such as smart energy systems, weather monitoring, and transport systems. The sampling rate can be very fine granularity, e.g., per second or millisecond. As a result, the sizes of data are typically large, along with diverse types and formats. Therefore, it is difficult to orchestrate smart city data sets.

In this paper, we present a private cloud platform, *SciCloud*, for assisting our smart city research project [6]. SciCloud as an ICT infrastructure is designed to handle different aspects related to smart cities. In particular, it deals with the data streams originating from smart systems, such as the energy systems and IoT networks that our project emphasizes. In order to facilitate the use of SciCloud, we have developed a scalable data framework to simplify smart city data management, including data collection, cleaning, anonymization, and publishing. This paper describes the applicability of the Cloud and the framework from the perspective of research, and demonstrates two use cases in the field of smart energy and air quality. The Cloud and the data management platform can be applied to many other use cases, due to their flexibility.

II. Related Works

Cloud computing becomes increasingly popular today, due to its ability of providing unlimited resources. Customers can use cloud resources based on the pay-as-you-go business model, which only pays for what actually have been used, e.g, the number of cores, the amount of memory and space. There are a number of well-known public cloud platforms available, including Google Cloud, Amazon EC2, Microsoft Azure, Rackspace and AliCloud. Many organizations intend to set up their own private cloud to get better privacy protection and to support their business goals. The presented SciCloud is a private cloud platform that provides a secure environment for our research of smart cities.

Clouds need effective tools to manage their computing resources. There are several main-stream open source cloud management frameworks, including OpenNebula [19], OpenStack [23], Eucalyptus [22], Nimbus [13], Snooze [9] and Origo [25] (used in the SciCloud). These frameworks greatly facilitate the cloud management: allocating computing resources according to user demands, failure recovery, load balancing, system monitoring and others. In order to provide a secure cloud computing platform, in this paper we present a novel smart city data management framework in this private cloud environment, and create an in-place data analytic service to support our scientific research of smart cities.

Smart city data are often characterized with the Big data characteristics: high volume, high variety and high velocity,

2378-3915/17 $31.00 © 2017 IEEE
DOI 10.1109/DEXA.2017.22

which is difficult to manage. Some attempts have been made towards smart city data management. Examples include the SCOPE [28], which is a cloud-based smart city open platform and ecosystem; CiDAP [5], which is a real-time smart city data platform; and FIWARE [10], which is a framework of providing intelligent application development in the Future Internet. They focus primarily on infrastructure development, data collection, test bench deployment or applications/services-specific development, but less emphasis on data sensitivity management. The [1], [4], [12] explore smart city data management from the perspectives of data security and privacy, which involve data collection, transmission, processing and visualization services. In contrast, in this paper we present a complete solution for simplifying smart city data management, which includes data extraction, processing, storing, sharing and publishing. Especially we emphasize privacy and sensitivity management of smart city data, and propose a three-level sensitivity model for publishing and sharing data.

III. The ICT infrastructure - SciCloud

The current trend of data processing is increasingly transferring traditional centralized solutions to cloud environments. Within our Centre of IT-Intelligent Energy System in Cities [6], we have created a private cloud platform to support our research in smart cities. The purpose of implementing a private cloud is to integrate infrastructure, platform and resources; and to provide a consolidated infrastructure for scientific experiments, software development and provision of analytics services.

Figure 1 illustrates the architecture of the SciCloud. It consists of 18 physical servers, with 80 cores and 564 GB of memory, and 4.2 TB of node storage, plus 1.2 TB of network storage.

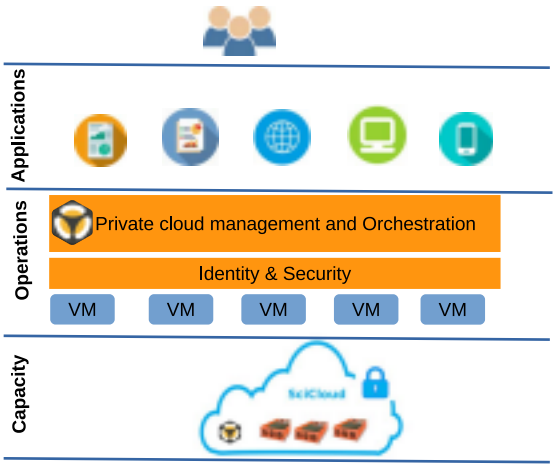

Fig. 1: The architecture of SciCloud

SciCloud is used to provide on-demand capacity for our research. SciCloud has integrated the Origo Virtual Infrastructure Engine [25] for resource management and to support a variety of demands of our researchers. Origo can efficiently manage virtual machines (VMs) that are scaled on a distributed infrastructure. Origo Virtual Infrastructure Engine provides the functionality to deploy, monitor and control VMs on a pool of distributed physical resources. Origo consists of the following three main components. The core is a centralized component that manages the lifecycle of a VM by performing basic VM operations, including deployment, monitoring, migration or termination. The core component also has a basic management and monitoring interface for the physical hosts. The second is the identity and security component. This component manages the security of user accounts and the safety of data in the cloud. The third is a capacity management component that adjusts the placement of VMs based on a set of predefined policies. The default capacity scheduler implements a simple pairing policy and supports user-driven integration constraints.

To support the use of this cloud, SciCloud offers Windows-based and Linux-based VM images with different pre-installed software packages. These include data science images with all the commonly used data analysis tools, such as R, Python, Pandas, Scikit-learn; and data management images with pre-installed different types of databases (e.g., PostgreSQL, MySQL, OpenTSDB, etc). Among others, many applications can be deployed in the VM. These settings can adequately satisfy the different needs of our research.

IV. Data Management Infrastructure

A. Smart city data management in the Cloud

Smart city data are collected from a variety of sources such as IoT devices, video surveillance systems, social networks, transport, government documents, or open data platforms, location-based services, and more. In addition, some data such as socio-economic data, contain sensitive personal related information such as social security number, name, age, home address and health, etc. Therefore, smart city data have the characteristics of big data, including big volume, high velocity and variety. These pose a grand difficulty in dealing with the data. In order to facilitate smart city data management, establishing a complete and flexible data management platform becomes essential. This is a key step between data sources and the applications of using the data. Although some studies have been done to study the big data platform of smart cities, most focus implementing a specific functional requirement and architectural design. At the same time, as there are many different tools and platforms with similar functionalities available in the big data community, we are often overwhelmed and confused by their features and capabilities. As a result, there is still a gap between a big data platform for smart cities and how they can be properly realized. We, therefore, present a cloud-based platform, called CITIESData [16], to manage smart city data, which is a trust framework that ensures that data can be properly shared, published, and used without compromising the data privacy.

The platform has the capability of processing the data with high diversity and complex, e.g., with different types, formats, meanings, and sizes; as well as handling the data with different quality issues, e.g., missing values and/or incorrect values. This platform aims at streamlining the whole data process: collecting, cleansing, storing, anonymizing, publishing and analyzing data. We have made a particular emphasis on data privacy and data quality. Figure 2 illustrates the system architecture which

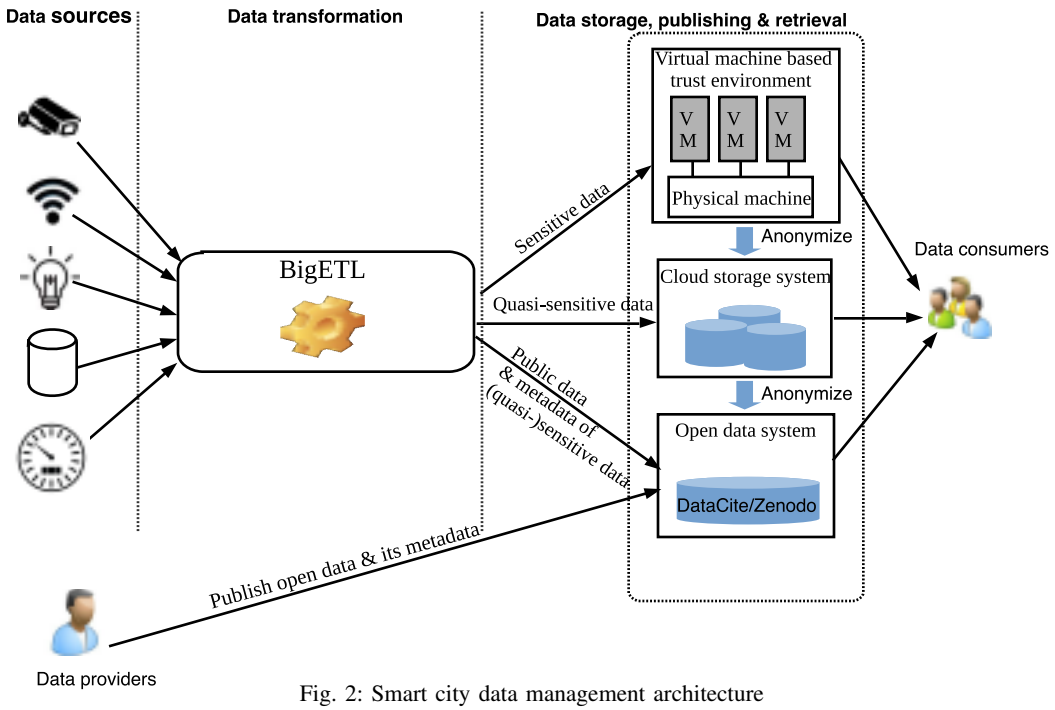

Fig. 2: Smart city data management architecture

comprises the phases including data extraction, transformation, anonymization, archiving and publishing.

This platform integrates a BigETL tool for data transformations (see https://github.com/xiufengliu/BigETL for more details), such as fixing missing values, removing duplicate data, and merging. The core feature of this system is on how to manage sensitive data for data publishing or sharing. This system classifies the data according to a three-level sensitivity model in terms of information disclosure. The data are classified into sensitive data, quasi-sensitive data, and public (open) data; and different strategies are used for the publishing. Sensitive data is shared with authorized users within a virtual machine based trust environment in the cloud - the data are not allowed to leave this environment when it is used. Quasi-sensitive data is shared via a cloud-based storage system, e.g., a private OwnCloud [26] that requires authorization; whereas public data is shared on an open data platform, namely Zenodo [29] or CKAN [8]. The sensitivity level can be mitigated by anonymization,e.g., sensitive data becomes non-sensitive after being anonymized. An open data platform itself is integrated with a data management system, such as CKAN, which allows data publishers to upload and publish data directly. An open data platform can also be restricted to publish metadata (i.e., the data of describing other data). This feature is useful for sharing information from (quasi-)sensitive data, i.e., only publishing the metadata while not the (quasi-)sensitive data itself. The benefit is that (quasi-)sensitive data can still be indexed and discovered through the open data platform even though the data themselves are not accessible. If a user needs to access (quasi-)sensitive data, (s)he has to link to a secure environment where user authorization is enforced. The open data portal is the single entrance to search the data available in all data repositories. This solution becomes a feasible way to maintain privacy and openness of smart city data.

In addition to the above "normal data life cycle", there is a demand for archiving research data, e.g., for a university archive and a national archive (see [16] for the details).

B. *In-place data analysis*

This data management system offers an additional setup for in-place analysis using data in the Cloud. This is done by adding a data analytics layer on top of CITIESData (see Figure 3). This layer has Jupyterhub and RStudio installed on the virtual machine, and both analytic tools can access data directly in the underlying CITIESData platform. As Jupyterhub and RStudio both offer the web-based interface for researchers to interact with the data, it means that they can do the analysis without copying the data out of the Cloud infrastructure. Another benefit is that the virtual machine has pre-installed all the necessary data analysis tools, software and packages. Researchers can be released from the tedious software installation, but focus on their analytics tasks. In addition, as virtual machines run on the Cloud, researchers can take full advantage of the computing power provided by the Cloud, for example, to handle a big data set which is usually not possible on a personal computer. The Jupyterhub and RStudio support multiple users, but each user has her/his own working environment. They can install additional software packages by themselves. Multiple users can share their work and work together, e.g., on the same notebook of Jupyter.

The data analysis platform itself is efficient, robust and secure (see [15], [17], [18] for more details). As mentioned

earlier, releasing data is an important service, and legislation requires the secure handling of sensitive data. To ensure handling of sensitive data safely, we design this architecture to store the data within the system. The data never leaves the secure area, and hence no data is copied to personal computers that may be backed up and propagated. In addition, this platform uses OwnCloud to assist data security management. In OwnCloud, a role-based model is applied to manage fine-level data permissions. For each user, only the data that (s)he should access will be granted read permission. The data will be automatically synchronized to the analytics layer for users to use. When logging into Jupyterhub or RStudio, (s)he can find the data residing in the home directory.

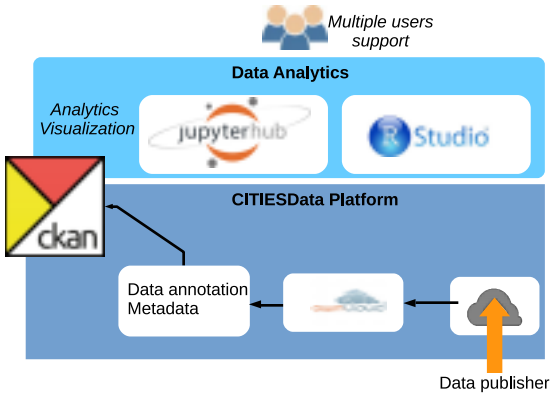

Fig. 3: In-place data analytics service

V. User Cases

In this section, we present two use cases as the examples of using the SciCloud platform.

A. Smart meter data management

A smart meter is an advanced meter that measures energy consumption at a regular time interval, typically every 15 minutes in electricity meters [15]. Smart meters communicate information back to the local utility for monitoring and billing purposes. The detailed energy usage information could lay bare the daily energy usage patterns of a household and even go so far as to enable deduction of what kind of device or appliance was in use at any given time. Besides, the unique ID of a smart meter is co-related to an individual or a household. This raises important privacy issues regarding the availability and processing of such data [27].

The data involved in our use case are the district heating data from 54 households in Sonderborg, Denmark. The meter readings are recorded every hour, and streamed into our system through the CITIESData platform. The heating consumption data is time-series data with two metrics, *volume* and *heat energy*, and the timestamp. The heating consumption data has some data quality issues, including row duplication and having missing values for some time series. Each time series has meter ID, which is associated with the household. The household data contains the sensitive information, including the name of customers and their addresses (road and building no.).

First, we address the data quality issues by cleansing the data, which involves fixing the missing values and outliers (anomaly high value above a set threshold value), transformation (extracting the date and hour from the timestamp), and removing the unnecessary values (e.g., the unit of the readings). This is done automatically by running a batch job in BigETL, which is triggered at a specified time.

Second, we address the data privacy problem by anonymizing the smart metering data so that information gleaned from it cannot easily be associated with an identified person. This is done by the following steps: 1) we replace the meter ID with a meaningless surrogate key, and decouple the time series from customer data; 2) we aggregate the time series, and provides daily and monthly readings to users; 3) we make use of our secure platform for ensuring the data safety. Data users are granted read permission in order to use the data. Data users can only do the online analytics within Jupyterhub or RStudio, while the data is not able to be distributed or downloaded onto personal computers.

Data quality and privacy protection are considered to be of prime importance to smart meter data management. This paper proposes a solution for streamlining smart meter data processing, and anonymizing high-frequency metering data through several strategies without compromising the use of the data. For these reasons, the SciCloud is well suitable for smart energy data management.

B. Air quality IoT data management

Air pollution is one of the most important factors that affects the health of people and the quality of life in smart cities. Over the past decades, the air quality in many global metropolitan cities has deteriorated significantly, especially in developing countries such as China and India. In the European countries, air pollution is not as prominent as the developing countries, but many governments have set their goals of reducing greenhouse gas emission, e.g, Danish government sets the goal of reducing greenhouse gas emission by 40% by 2020 and becoming a completely fossil-fuel free country by 2050. As a result, it is important to monitoring the air quality so as to provide the real-time information for the government and citizens for decision-making purposes. To support this, we have developed a cloud-based monitoring system [2], and deployed in Vejle, Denmark and Trondheim, Norway to monitor the air quality. We make use of the SciCloud in this project, and do the following:

- *Stream IoT data*: IoT sensors are installed in several places around the city, e.g., intersections of the roads, to monitor the air quality (including CO2 and CO level, particle sizes of PM1.0, 2.5 and 10) as well as weather condition (temperature, humidity, air pressure, wind speed). The resulting IoT data is streamed into the SciCloud in a real-time fashion. Besides, we also stream the traffic data at the places where the sensors are located into the SciCloud from a third-party traffic monitoring system.
- *Data storage*: To integrate multiple data sources, we deploy the time-series database, OpenTSDB [24], in the SciCloud to manage all the time series. All the

time series are stored in a uniform format, i.e., metric, timestamp, values, and a number of tags of labeling time series (e.g, the locations of the sensors). OpenTSDB is a distributed database, which can store and query data efficiently through its RESTful APIs. The data is provided as a service for application developers.

- *Air quality analytics*: There is a processing and data mining module, which is implemented as the Python program in Jupyter. The program reads the time series of air quality, weather condition and traffic, and studies impact of weather and traffic flow on the air quality by correlation. This model is done offline, but updates the analytic charts regularly through reading the data from OpenTSDB.
- *Monitoring dashboard*: Local city authorities are enabled to view air quality and the analytic results through a dashboard. The dashboard is implemented on the Apache Zeppelin [3] that is deployed on the SciCloud. The dashboard can be exported as an iFrame to be embedded into any websites, e.g., the government websites.

VI. Conclusion and Future Work

Smart cities necessitate management of the so-called Big Data. In this paper, we have presented a private cloud platform, SciCloud, for smart city data management. SciCloud provides the on-demand computing resources for researchers to experiment, develop prototypes and proof of concept. To facilitate the use of this cloud, we have developed a secured scalable smart data management system to streamline the processing of the data life cycle. In addition, we have proposed an in-place analytics environment for analyzing the data in the Cloud, and we verified its effectiveness by demonstrating two use cases on this Cloud platform.

In future work, we intend to further develop this SciCloud to support other research projects, and verify the Cloud by managing the data from more domains.

Acknowledgement

This research was supported by the CITIES project (NO. 1035-00027B) funded by Innovation Fund Denmark. The infrastructure components are partly supported by the Danish Electronic Infrastructure (DeIC) through the project "Science Cloud for Cities".

References

[1] A.G. Abbasi, Z.A. Khan, and Z. Pervez, "Towards Cloud Based Smart Cities Data Security and Privacy Management," Proc. of the 7th International Conference on Utility and Cloud Computing, pp. 806–811, 2014.

[2] D. Ahlers, P.A. Driscoll, K.F. Kraemer, F.V. Anthonisen, and J. Krogstie, "A Measurement-Driven Approach to Understand Urban Greenhouse Gas Emissions in Nordic Cities," NIK, 2016.

[3] Apache Zeppelin. Avail. at http://zeppelin.apache.org/ as of 2017-5-15.

[4] J. Bohli, D. Garcia, P. Langendorfer, M.V. Moreno, and A. Skarmeta, "SMARTIE project: Secure IoT data management for smart cities," Proc. of the International Conference on Recent Advances in Internet of Things (RIoT), pp. 1–6, 2015.

[5] B. Cheng, S. Longo, F. Cirillo, M. Bauer, and E. Kovacs, "Building a big data platform for smart cities: Experience and lessons from santander," In Big Data (BigData Congress), pp. 592–599, 2015.

[6] CITIES. Avail. at http://smart-cities-centre.org as of 2017-5-15.

[7] CityZenith. Avail. at http://cityzenith.com as of 2017-5-15.

[8] CKAN: The open source data portal software. Avail. at https://ckan.org as of 2017-5-15.

[9] E. Feller, C. Morin, and L. Rilling, "Snooze: A Scalable and Autonomic Virtual Machine Management Framework for Private Clouds," Proc. of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 482–489, 2012.

[10] FIWARE: Open Source Platform. Avail. at http://www.fi-ware.org as of 2017-5-15.

[11] J. Gantz, and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," IDC iView: IDC Analyze the Future, vol. 2007, pp. 1–16, 2012.

[12] N. Gruschka, and D. Gessner, "IoT-A: Internet of Things Architecture, Project Deliverable D4.2 - Concepts and Solutions for Privacy and Security in the Resolution Infrastructure," 2012.

[13] K. Keahey, T. Freeman, J. Lauret, and D. Olson. "Virtual workspaces for scientific applications," Journal of Physics: Conference Series, 78(1), 2007.

[14] S.A. Kim, D. Shin, Y. Choe, T. Seibert, and S.P. Walz, "Integrated energy monitoring and visualization system for Smart Green City development: Designing a spatial information integrated energy monitoring model in the context of massive data management on a web based platform," Autom Constr, 22:51–9, 2012.

[15] X. Liu, and P.S. Nielsen, "A hybrid ICT-solution for smart meter data analytics," Energy, 115(3):1710–1722, 2016.

[16] X. Liu, A. Heller, and P.S. Nielsen, "CITIESData: A Smart City Data Management Framework," Knowledge and Information Systems. 2017:1-24, 2017.

[17] X. Liu, L. Golab, W. Golab, I.F. Ilyas, and S. Jin, "Smart Meter Data Analytics: Systems, Algorithms, and Benchmarking," 42:139, 2016.

[18] X. Liu, and P.S. Nielsen, "Streamlining Smart Meter Data Analytics," Proc. of the 10th Conference on Sustainable Development of Energy, Water and Environment Systems, SDEWES2015.0558,1-14, 2015.

[19] D. Milojicic, I.M. Llorente, and R.S. Montero, "OpenNebula: A cloud management tool". IEEE Internet Computing, vol. 15, March 2011.

[20] P. Neirotti, A. De Marco, A.C. Cagliano, G. Mangano, and F. Scorrano, "Current trends in Smart City initiatives: Some stylised facts,". Cities 2014; 38:25–36.

[21] New Stats From The State Of Cloud Report. Avail. at https://www.forbes.com/sites/ benkepes/2015/03/04/new-stats-from-the-state-of-cloud-report/ as of 2017-5-15.

[22] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, "The Eucalyptus open-source cloud-computing system". Proc. of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009.

[23] OpenStack: Open source cloud computing software, Avail. at http://www.openstack.org as of 2017-5-15.

[24] OpenTSDB. Avail. at http://OpenTSDB.net as of 2017-05-15.

[25] Origo: cloud computing. Avail. at https://www.origo.io as of 2017-5-15.

[26] Owncloud. Avail. at http://owncloud.org as of 2017-05-15.

[27] E.L. Quinn, "Privacy and the New Energy Infrastructure," Social Science Research Network (SSRN), 2009.

[28] SCOPE: A Smart-city Cloud-based Open Platform and Ecosystem. Avail. at http://www.bu.edu/hic/research/scope/ as of 2017-5-15.

[29] Zenodo. Avail. at http://zenodo.org as of 2017-05-15.